DOCUMENT RESUME

ED 109 148 TM 004 588

AUTHOR Brennan, Robert L.; Kane, Michael F.
TITLE The Generalizability of Class Means.
PUB DATE [Apr 75]
NOTE 54p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$3.32 PLUS POSTAGE
DESCRIPTORS Analysis of Variance; *Classes (Groups of Students); Correlation; Error Patterns; Item Analysis; Measurement Techniques; *Scores; Statistical Analysis; *Test Interpretation; *Test Reliability
IDENTIFIERS Class Means; *Generalizability Theory

ABSTRACT

When classes are the units of analyses, estimates of the reliability of class means are needed. Using classical test theory it is difficult to treat this problem adequately. Generalizability theory, however, provides a natural framework for dealing with the problem. Each of four possible formulas for the generalizability of class means is derived from two points of view. Each of the four generalizability coefficients is shown to be the expected value of a variation on the split-half method for estimating reliability. Finally, the four coefficients are related to estimates of the reliability of class means that have been used previously. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED109148

TM 004 588

Session 19:20

The Generalizability of Class Means

Robert L. Brennan and Michael T. Kane

State University of New York at Stony Brook

Department of Education

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Running head: The Generalizability of Class Means

Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975.

Biographical Resume of Authors

Brennan, Robert L. Position: Assistant Professor of Education. Address: Department of Education, SUNY at Stony Brook, Stony Brook, New York 11794. Degrees: B.A. State College at Salem, Mass., M.A.T., Ed.D. Harvard University. Specialization: Psychometrics, statistics, and evaluation. AERA divisional membership: B, C, and D.

Kane, Michael T. Position: Assistant Professor of Education. Address: Department of Education, SUNY at Stony Brook, Stony Brook, New York 11794. Degrees: B.S. Manhattan College, M.A. State University of New York at Stony Brook, M.S., Ph.D. Stanford University. Specialization: Psychometrics and statistics. AERA divisional membership: D.

Abstract

When class is the unit of analysis, estimates of the dependability of class means are frequently required. Using classical test theory it is difficult to treat this problem adequately. In this paper, we consider the dependability of class means by applying generalizability theory to a split-plot design in which students are nested within classes. Using the split-plot design we obtain four distinct generalizability coefficients. We then compare these four coefficients with each other and with three previously reported reliability coefficients. We find that each of the three reliability coefficients is related to one or more generalizability coefficients. However, none of the reliability coefficients is equivalent to the generalizability coefficient which is, in our judgment, usually the most appropriate coefficient for describing the dependability of class means.

The Generalizability of Class Means

Introduction

Recently, a number of researchers have given serious consideration to the problem of estimating reliability when the unit of analysis is a class mean or some other aggregate score for a set of persons. Haney (1974a, 1974b) has reviewed important aspects of the relevant literature.

The study of this topic has been motivated by the analysis of data from several different sources. In particular, large scale evaluations, such as those undertaken for Head Start (see Smith & Bissell, 1970) and Follow Through (see Abt Associates, 1974), frequently require estimates of reliability when class is the unit of analysis. Similar issues arise in the study of course evaluation questionnaires (see Kane, Gillmore, & Crooks, 1974).

Using concepts from classical reliability theory, Shaycroft (1962), Wiley (1970), and Thrash and Porter (1974) have developed three different coefficients for estimating the reliability of class means. The procedures used to develop these three coefficients all assume that an observed score is the sum of a true score and an undifferentiated error term. However, each of these procedures makes different specific assumptions about what constitutes an appropriate

estimate of the error variance. As a result, each procedure gives a different estimate of the reliability of class means. Since the three procedures will not, in general, lead to even approximately equal estimates of the reliability of class means, it is of considerable importance to determine the appropriate coefficient for any particular application.

Within the context of classical reliability theory, it is difficult to compare the three procedures and arrive at reasonable conclusions about their relative merits. Difficulty in comparing the proposed reliability coefficients exists because these coefficients are derived from statistical models which are based on different assumptions. In particular, the variance attributed to error arises from different sources in the three models. However, the coefficients can be compared directly within the context of a more comprehensive and detailed model.

Brennan (in press), Kane et al. (1974), and Haney (1974a) have suggested that the reliability of class means be approached through generalizability theory, as explicated by Cronbach, Gleser, Nanda, and Rajaratnam (1972). Generalizability theory extends reliability theory by allowing for a multi-dimensional interpretation of error which, in turn, provides a much more systematic and precise method for studying the dependability of class means.

In this paper, we consider the dependability of class means by applying generalizability theory to a split-plot design in which students are nested within classes. Using the split-plot design we obtain four distinct generalizability coefficients. We then compare these four generalizability coefficients with each other and with the three previously reported reliability coefficients. We find that each of the three reliability coefficients is related to one or more generalizability coefficients. However, none of the reliability coefficients is equivalent to the generalizability coefficient which is, in our judgment, usually the most appropriate coefficient for describing the dependability of class means.

Generalizability Theory

For a thorough presentation of generalizability theory, see The Dependability of Behavioral Measurements (Cronbach et al., 1972). A briefer introduction to many of the basic ideas is found in Lindquist (1953). Here we discuss some concepts from generalizability theory that relate to our subsequent treatment of the dependability of class means.

Overview

The purpose of both generalizability theory and reliability theory is to characterize the dependability of measurements. Classical reliability theory assumes that errors of measurement are sampled from an undifferentiated univariate distribution. In order to estimate the proportion of observed score variance attributable to error, reliability theory uses correlations and one-way analysis of variance (ANOVA). By contrast, generalizability theory recognizes the existence of multiple sources of error, and allows for the use of any ANOVA design in order to estimate the magnitude of the variance components.

Although generalizability theory borrows its statistical models and research designs from ANOVA, there are some changes in terminology and interpretation. ANOVA is typically used to test the statistical significance of hypotheses; the mean squares

or estimated components of variance are not of primary interest. In generalizability theory the emphasis is reversed. The components of variance and the coefficients of generalizability computed from these variance components are interpreted as descriptive statistics; statistical significance plays no essential role. Cronbach et al. (1972, p. 192) actually advise against the use of tests of significance for the variance components.

Terminology

In generalizability theory, any observation on some unit of analysis (e.g., school, class, or student) is taken from a larger set or universe of observations. Any observation from this universe can be characterized by the conditions under which it is made. The set of all possible conditions of a particular kind is called a facet. For example, when class is the unit of analysis, the conditions of observation are characterized by an item facet and a student facet. This terminology is slightly different from that typically used in statistics, where classes, items, and students would all be referred to as dimensions. The use of the term "facet" in generalizability theory serves to emphasize the distinction between the unit of analysis which is being observed and the facets, which indicate the conditions under which the observations are made.

Generalizability theory also emphasizes the distinction between G studies, which examine the dependability of some measurement procedure, and D studies, which provide the data for substantive decisions. The purpose of the G study is to estimate components of variance, which may then be used to estimate generalizability coefficients for a variety of D studies. The G study and the D study may be the same study, or they may be different studies using the same design. Generally, however, G studies are most useful when they employ complex designs and large sample sizes to provide stable estimates of as many variance components as possible. These components can then be used to estimate generalizability coefficients for various D study designs before any D study is implemented.

In the discussion that follows, it will be useful to employ some additional terminology from generalizability theory. According to Cronbach et al (1972),

> (t)he test developer or other investigator who carries out a G study takes certain facets into consideration and, with respect to each facet, considers a certain range of conditions. The observations encompassed by the possible combinations of conditions that the G study represents is called the universe of admissible observations. We may also speak of the universe of admissible conditions of a certain facet.

> A decision maker, applying essentially the same measuring technique, proposes to generalize to some universe of conditions all of which he sees as eliciting samples of the same information. We refer to that as the universe of generalization. The G study can serve this decision maker only if its universe of admissible conditions is identical to or includes the proposed universe of generalization. Different decision makers may propose different universes of generalization. A G study that defines the universe of admissible observations broadly, encompassing all the likely universes of generalization, will be useful to various decision makers. (p. 20)

The decision maker is generally interested in the mean of the observations over the universe of generalization, called the universe score. Universe scores are not directly observable and are usually estimated by the mean over some sample of observations. The dependability of these estimates is reflected by the value of a generalizability coefficient.

Coefficients of generalizability are defined as the ratio of the universe score variance to the expected observed score variance. These coefficients are essentially intraclass correlation coefficients with universe score variance replacing the true score variance of classical test theory (Cronbach,

Ikeda, & Avner, 1964). The observed score variance has essentially the same interpretation in both formulations.

Background

Many of the ideas that underlie generalizability theory are not new. In fact, Lindquist (1953) and, to a lesser extent, Hoyt (1941) suggested procedures for calculating reliability that foreshadow the approach of generalizability theory. Using a randomized block design and basic principles from reliability and analysis of variance, both Lindquist and Hoyt calculated intraclass correlation coefficients (generalizability coefficients) that are algebraically equivalent to Kuder and Richardson's (1937) Formula 20 and Cronbach's (1951) Coefficient α (see Brennan, in press).

Generalizability theory extends the work of Hoyt and Lindquist to multi-faceted statistical and measurement models. In subsequent sections of this paper we concentrate upon one such model, the split-plot design, as a basis for considering the dependability of class means.

The Split-Plot Design as a Model for the Generalizability of Class Means

Split-Plot Design

In most situations, when the dependability of class means is under study, the design is that designated by Cronbach et al. (1972) as design V-B. This design is often referred to as a

split-plot design in standard experimental design texts (e.g., Kirk, 1968, and Winer, 1971). In this design students are nested within classes and crossed with items. Thus, each class contains a different set of students, but the same set of items are administered to the students in all classes.

The structural model for this design is:

$$X_{\underline{csi}} = \mu + \alpha_{\underline{c}} + \pi_{\underline{s}(\underline{c})} + \beta_{\underline{i}} + \alpha\beta_{\underline{ci}} + \beta\pi_{\underline{is}(\underline{c})} + e_{\underline{o}(\underline{csi})} , \quad (1)$$

where

μ = grand mean,

$\alpha_{\underline{c}}$ = effect for class $\underline{c}$ ($\underline{c}$ = 1, 2, ... , $\underline{n_c}$),

$\pi_{\underline{s}(\underline{c})}$ = effect for student $\underline{s}$ ($\underline{s}$ = 1, 2, ... , $\underline{n_s}$) nested within class $\underline{c}$,

$\beta_{\underline{i}}$ = effect for item $\underline{i}$ ($\underline{i}$ = 1, 2, ... , $\underline{n_i}$),

$\alpha\beta_{\underline{ci}}$ = class by item interaction,

$\beta\pi_{\underline{is}(\underline{c})}$ = item by person (nested within class) interaction, and

$e_{\underline{o}(\underline{csi})}$ = experimental error ($\underline{o}$ is a replication subscript).

Following Kirk (1968), the subscript $\underline{c}$ referring to the nested treatment, class, is placed within parentheses. To simplify our discussion, we will assume in this paper that the number of students within a class, $\underline{n_s}$, is a constant over all classes.

Factorial Design

The nesting of students within classes indicated in Equation 1 results in a confounding of the effects due to classes and students. The implications of this confounding in the split-plot design are evident from a consideration of the analogous three-way factorial design in which classes, students, and items are all crossed:

$$X_{csi} = \mu + \alpha_c + \pi_s + \alpha\pi_{cs} + \beta_i + \alpha\beta_{ci} + \beta\pi_{is} + \alpha\beta\pi_{cis} + e_{o(csi)}. \quad (2)$$

If the factorial design were appropriate, then every student would appear in each and every class, and the student effect could be estimated independently of the class effect.

Confounding in the Split-Plot Design

The differences between the model equations for the split-plot and factorial designs are attributable to the fact that in the factorial design (Equation 2) students are crossed with classes, and in the split-plot design (Equation 1) students are nested within classes. This nesting in the split-plot design results in a confounding of at least two sets of effects represented in Equation 2.

First, the student main effect is confounded with the class by student interaction; that is, $\pi_{s(c)}$ in the split-plot design

represents the combined contribution of the student effect and the student by class interaction, $\pi_{\underline{s}} + \alpha\pi_{\underline{cs}}$, in the factorial design. For the split-plot design, each student is observed in only one class; thus, there is no way to estimate the student effect independent of the class by student interaction.

Similarly, the student by item interaction is confounded with the class by student by item interaction; thus, $\beta\pi_{\underline{is}(\underline{c})}$ in the split-plot design takes the place of $\beta\pi_{\underline{is}} + \alpha\beta\pi_{\underline{cis}}$ in the factorial design.

Also, in the absence of replications, there is another type of confounding in both designs. When there is only one observation for a given student responding to a given item, then, for Equation 1, the error term, $e_{\underline{o}(\underline{csi})}$, is confounded with the item by student (nested within class) interaction, $\beta\pi_{\underline{is}(\underline{c})}$. In this case, $\beta\pi_{\underline{is}(\underline{c})}$ in the split-plot design replaces $\beta\pi_{\underline{is}} + \alpha\beta\pi_{\underline{cis}} + e_{\underline{o}(\underline{csi})}$ in the factorial design.

Terminology

The terminology used in the preceding section is, for the most part, representative of the terminology used in texts on experimental design (e.g. Kirk, 1968, and Winer, 1971). Since class is the unit of analysis under consideration here, Cronbach et al. (1972) would say that the split-plot and factorial designs have two facets (students and items).

Furthermore, Cronbach et al. (1972) would distinguish between the G study design used to estimate variance components and the D study design used to make decisions. In theory, the G study design and the D study design need not be the same; however, for purposes of simplicity, unless otherwise noted we assume here that both the G and D studies employ a split-plot design.

Assumptions

The assumptions for the split-plot design model in Equation 1 are well documented in the literature and experimental design texts. However, we wish to emphasize two of these assumptions. First, each effect in the model is assumed to be independent of every other effect. Second, in order to make the estimates of the effects unique, the expected value of each effect over any of its subscripts is set equal to zero.

This second assumption is especially critical to an understanding of subsequent parts of this paper; however, this assumption is easily misunderstood. Consider the effect $\alpha_{\underline{c}}$ in Equation 1. Suppose, for some study, we take a sample of $\underline{n_c}$ classes from a population of $\underline{N_c}$ classes. The second assumption implies that the sum of $\alpha_{\underline{c}}$ over the population of $\underline{N_c}$ classes is constrained to be zero, and the sum of the estimates of $\alpha_{\underline{c}}$ over the sample of $\underline{n_c}$ classes is constrained to be zero. However, it is not necessarily true that the sum of $\alpha_{\underline{c}}$ over the sample of $\underline{n_c}$

classes is zero. Using "·" to indicate a sample mean, "-" to indicate a population mean, and "^" to indicate the estimated value, the second assumption means that $\alpha_{\cdot}$ equals zero, $\hat{\alpha}_{\cdot}$ equals zero, but $\alpha_{\cdot}$ does not necessarily equal zero.

Generalizability Coefficients from a Split-Plot Design

By definition, a generalizability coefficient is the ratio of the universe score variance to the expected value of the observed score variance. For the split-plot model, with class as the unit of analysis, four different generalizability coefficients can be obtained. Each of these coefficients is characterized by a different definition of universe score and, hence, a different definition of error. However, for each of these coefficients, the expected observed score variance is identical.

Expected Observed Score Variance

Because the effects in Equation 1 are assumed to be sampled independently, the expected observed score variance is simply the sum of the variances of the separate effects. For a sample of n_i items and n_s students, the expected observed score variance for class means is:

$$\mathcal{E}\sigma^2(X_{\underline{c}\cdot\cdot}) = \sigma^2(\alpha) + \frac{\sigma^2(\pi,\alpha\pi)}{n_{\underline{s}}} + \frac{\sigma^2(\alpha\beta)}{n_{\underline{i}}} + \frac{\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e})}{n_{\underline{i}}n_{\underline{s}}}, \quad (3)$$

where $\sigma^2(\pi,\alpha\pi)$ is the population variance of $\pi_{\underline{s}(\underline{c})}$ in Equation 1. The notation $\sigma^2(\pi,\alpha\pi)$ is used to emphasize that the student effect, $\pi_{\underline{s}}$, and the class by student interaction, $\alpha\pi_{\underline{cs}}$, are confounded in the effect $\pi_{\underline{s}(\underline{c})}$. A similar interpretation can be given to $\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e})$ assuming there is only one observation for a given student responding to a given item. There is no variance component for the item effect because the item effect is constant for all classes, and, therefore, $\sigma^2(\beta)$ is zero. Note that the population variance components in Equation 3 are for samples of one item and one student.[1]

In subsequent sections we develop the universe score variance and the associated generalizability coefficient for four different universes of generalization: (a) an infinite universe of students and items, (b) an infinite universe of students and a finite set of items, (c) a finite set of students and an infinite universe of items, and (d) a finite set of students and a finite set of items.

Infinite Universe of Students and Items

If the objective of a D study is to generalize to an infinite universe of students and items, then the universe of generalization is the completely crossed universe of

admissible observations. This analysis treats the set of items used in the D study as a sample from an infinite universe of items that could have been used to measure the general outcome that is of interest; and the analysis treats the students in each class as a sample from the infinite universe of students who might have been included in the class. The universe score for each class is the expected value of the observed mean over all possible samples of students and items in the universe of admissible observations, and is given by

$$\nu_{\underline{c}} = \mu + \alpha_{\underline{c}} \ . \qquad (4)$$

The corresponding universe score variance is

$$\sigma^2(\nu_{\underline{c}}) = \sigma^2(\alpha_{\underline{c}}) \ . \qquad (5)$$

The universe score can also be obtained by taking the limit of the observed score variance in Equation 3 as $\underline{n}_{\underline{i}}$ and $\underline{n}_{\underline{s}}$ both go to infinity.

For generalization to an infinite universe of students and items, the appropriate generalizability coefficient is the ratio of universe score variance (Equation 5) to expected observed score variance (Equation 3):

$$\mathcal{E}\rho^2(\underline{S},\underline{I}) = \frac{\sigma^2(\alpha)}{\sigma^2(\alpha) + \left[\frac{\sigma^2(\pi,\alpha\pi)}{n_{\underline{s}}} + \frac{\sigma^2(\alpha\beta)}{n_{\underline{i}}} + \frac{\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e})}{n_{\underline{i}}n_{\underline{s}}}\right]} . \quad (6)$$

The notation $\mathcal{E}\rho^2$ is consistent with Cronbach et al. (1972) and indicative of the fact that generalizability coefficients can be interpreted as squared correlation or intraclass correlation coefficients. The letters $\underline{S}$ and $\underline{I}$ in parentheses indicate that the universe of generalization is an infinite universe of students ($\underline{S}$) and an infinite universe of items ($\underline{I}$). The brackets in the denominator identify the components of variance that jointly constitute error variance when the universe score variance is $\sigma^2(\alpha)$. Thus, the expected observed score variance can be viewed as universe score variance plus error variance.

Equation 6 results from the assumption that the universe of generalization is, in effect, the completely crossed universe of admissible observations. The substantive questions in a particular D study may indicate that the universe of generalization should be restricted to some subset of the completely crossed universe of admissible observations. That is, the investigator may wish to generalize to some finite subset of the possible conditions (or levels) for one or more facets in the universe of admissible observations. In particular, in

some cases, it may not be appropriate to generalize beyond the specific conditions of some facet(s) included in the D study.

For example, an investigator who is interested in how well a training program has taught students to perform a particular set of mechanical tasks is not interested in generalizing to a broad set of such tasks. The finite set of tasks on which observations are taken may constutute the universe of generalization for the task facet. (By contrast, if the hypothesis under consideration concerns general mechanical ability, the universe of generalization would be taken as an infinite universe of possible tasks that might have been observed.)

Theoretically, then, for the split-plot design, some D studies may require a universe of generalization in which the set of items (or tasks) is finite, the set of students is finite, or both are finite.

Infinite Universe of Students; Finite Set of Items

If generalization is to the finite set of items included in the D study, then the universe score is the expected value of the observed mean score ($X_{\underline{c}\cdot\cdot}$) over that particular set of items and over all students. The components that enter into the observed score are not changed by restricting the universe of generalization, and Equation 1 is still the appropriate model. In taking the expected value of $X_{\underline{c}\cdot\cdot}$, only terms with $\underline{s}$ as a subscript become zero, and the universer score is

given by

$$\nu_{\underline{c}} = \mu + \alpha_{\underline{c}} + \beta_{\cdot} + \alpha\beta_{\underline{c}\cdot} \quad . \tag{7}$$

The item effect and the class by item interaction are present in Equation 7 because the expected value is <u>not</u> taken over all items in the universe of admissible observations, and there will, in general, be systematic effects due to the finite set of items included in the universe of generalization.

The universe score variance corresponding to Equation 7 is

$$\sigma^2(\nu_{\underline{c}}) = \sigma^2(\alpha) + \frac{\sigma^2(\alpha\beta)}{\underline{n}_{\underline{i}}} \quad . \tag{8}$$

Equation 8 can also be derived from Equation 3 by taking the limit as $\underline{n}_{\underline{s}}$ approaches infinity. Again, $\beta_{\cdot}$ is a constant for all classes, and $\sigma^2(\beta)$ is zero.

For generalization to an infinite universe of students ($\underline{S}$) and the finite set of items ($\underline{I}^*$) used in the D study, the generalizability coefficient is obtained from Equations 3 and 8:

$$\mathcal{E}\rho^2(\underline{S},\underline{I}^*) = \frac{\sigma^2(\alpha) + \dfrac{\sigma^2(\alpha\beta)}{\underline{n}_{\underline{i}}}}{\left[\sigma^2(\alpha) + \dfrac{\sigma^2(\alpha\beta)}{\underline{n}_{\underline{i}}}\right] + \left[\dfrac{\sigma^2(\pi,\alpha\pi)}{\underline{n}_{\underline{s}}} + \dfrac{\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e})}{\underline{n}_{\underline{i}}\underline{n}_{\underline{s}}}\right]} \quad . \tag{9}$$

Infinite Universe of Items; Finite Set of Students

In educational research and evaluation, it is generally inappropriate to restrict the universe of generalization for the student facet. For diagnostic purposes, we may be interested in the universe score for a single student, but class means are seldom used in this way. In program evaluation and research, the intention is almost always to generalize to some population of present and/or future students.

Nevertheless, one can obtain the generalizability coefficient for a finite set of students and an infinite universe of items. Later we will show that this coefficient corresponds to one of the statistics reported in the literature for estimating the reliability of class means. For this universe of generalization, the universe score is

$$\nu_{\underline{c}} = \mu + \alpha_{\underline{c}} + \pi_{\cdot(\underline{c})} \ , \tag{10}$$

and the universe score variance is

$$\sigma^2(\nu_{\underline{c}}) = \sigma^2(\alpha) + \frac{\sigma^2(\pi,\alpha\pi)}{n_{\underline{s}}} \ . \tag{11}$$

The variance due to the student effect does not go to zero because a different set of students is in each class. Equation 11 can also be derived from Equation 3 by taking the limit as $n_{\underline{i}}$ approaches infinity.

For generalization to an infinite universe of items ($\underline{I}$) and the finite set of students ($\underline{S}^*$) used in the D study, the generalizability coefficient is obtained from Equations 3 and 11:

$$\mathcal{E}\rho^2(\underline{S}^*,\underline{I}) = \frac{\sigma^2(\alpha) + \dfrac{\sigma^2(\pi,\alpha\pi)}{\underline{n}_{\underline{s}}}}{\left[\sigma^2(\alpha) + \dfrac{\sigma^2(\pi,\alpha\pi)}{\underline{n}_{\underline{s}}}\right] + \left[\dfrac{\sigma^2(\alpha\beta)}{\underline{n}_{\underline{i}}} + \dfrac{\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e})}{\underline{n}_{\underline{i}}\underline{n}_{\underline{s}}}\right]} . \quad (12)$$

Finite Set of Students; Finite Set of Items

Restricting generalization to a particular set of students and a particular set of items is even less likely to be appropriate than restricting the universe for either facet and generalizing over the other. The results are presented here because they lead to a coefficient that corresponds to a reliability coefficient that has been proposed for class means. The universe score for a fixed set of students in each class, and a fixed set of items for all classes is

$$\nu_{\underline{c}} = \mu + \alpha_{\underline{c}} + \pi_{\cdot(\underline{c})} + \beta_{\cdot} + \alpha\beta_{\underline{c}\cdot} + \beta\pi_{\cdot\cdot(\underline{c})} ; \quad (13)$$

and the universe score variance is

$$\sigma^2(\nu_{\underline{c}}) = \sigma^2(\alpha) + \frac{\sigma^2(\pi,\alpha\pi)}{\underline{n}_{\underline{s}}} + \frac{\sigma^2(\alpha\beta)}{\underline{n}_{\underline{i}}} + \frac{\sigma^2(\beta\pi,\alpha\beta\pi)}{\underline{n}_{\underline{i}}\underline{n}_{\underline{s}}} . \quad (14)$$

The universe score variance is estimable if the effects $\beta\pi_{\underline{is}(\underline{c})}$ and $\underline{e}_{\underline{o}(\underline{csi})}$ in Equation 1 are not confounded; that is,

Equation (14) is estimable if there is more than one replication of each class-student-item observation. This coefficient can also be estimated if it can be assumed that $\sigma^2(\beta\pi,\alpha\beta\pi)/\underline{n}_i\underline{n}_s$ equals zero; in this case, the true score variance is given by the first three terms in Equation 14, and these variance components are all estimable.

For generalization to the finite set of items ($\underline{I}^*$) and the finite set of students ($\underline{S}^*$) in the D study, the generalizability coefficient is obtained from equations 3 and 14:

$$\mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*) = \frac{\sigma^2(\alpha) + \dfrac{\sigma^2(\pi,\alpha\pi)}{\underline{n}_s} + \dfrac{\sigma^2(\alpha\beta)}{\underline{n}_i} + \dfrac{\sigma^2(\beta\pi,\alpha\beta\pi)}{\underline{n}_i\underline{n}_s}}{\left[\sigma^2(\alpha) + \dfrac{\sigma^2(\pi,\alpha\pi)}{\underline{n}_s} + \dfrac{\sigma^2(\alpha\beta)}{\underline{n}_i} + \dfrac{\sigma^2(\beta\pi,\alpha\beta\pi)}{\underline{n}_i\underline{n}_s}\right] + \dfrac{\sigma^2(\underline{e})}{\underline{n}_o\underline{n}_i\underline{n}_s}}, \quad (15)$$

where $\underline{n}_o$ is the number of replications of each class-student-item observation.

Classical Reliability and the Spearman-Brown Correction

All four of these generalizability coefficients have the general form of a reliability coefficient if true score is defined to be equal to the appropriate universe score. The differences among the coefficients are, then, the differences among their definitions of true score and error score.

For $\mathcal{E}\rho^2(\underline{S},\underline{I})$, Equation 6, the universe score variance is the sampling variance of the main effect due to classes, $\sigma^2(\alpha)$.

All other components in the observed score variance are sources of error. In classical test theory, the error variance is undifferentiated; increasing the number of observations by a factor of $\underline{M}$ leaves the true score variance unchanged and decreases the error variance by $1/\underline{M}$. This regularity is the basis for the Spearman-Brown formula for changes in the length of a test. In Equation 6, the error variance has no such simple relationship to the number of students, items, or the product of the two; consequently, the Spearman-Brown formula does not apply. It is, however, easy to compute $\mathcal{E}\rho^2(\underline{S},\underline{I})$ for any number of students and items by substituting the appropriate values of $\underline{n}_{\underline{s}}$ and $\underline{n}_{\underline{i}}$ in Equation 6.

For $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$, Equation 9, where interest is restricted to the finite set of items in the D study, the class by item interaction is a component in the true score variance. For this coefficient, increasing the number of students by a factor of $\underline{M}$ will decrease the error variance by $1/\underline{M}$ but will not affect the universe score variance. Thus, the Spearman-Brown formula holds for the number of students. However, increasing the number of items by a factor of $\underline{M}$ does not decrease the error variance by $1/\underline{M}$ and does affect the universe score variance. Thus, the Spearman-Brown formula does <u>not</u> hold for items.

Similarly, the Spearman-Brown formula applies to $\mathcal{E}\rho^2(\underline{S}^*,\underline{I})$ for changes in the number of items but not for changes in the number of students. Finally, the Spearman-Brown formula applies to $\mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*)$ for changes in the number of replications,

but does not apply for changes in the number of students or the number of items, because such changes affect the universe score variance.

Estimation of Variance Components

The process of obtaining numerical estimates of generalizability coefficients usually involves two steps. First, the components of variance are estimated from the G study. Then, the generalizability coefficient is calculated using the estimated variance components and the sample sizes from the D study.

General procedures for the estimation of variance components from computed mean squares are discussed by Cornfield and Tukey (1956), Cronbach et al. (1972), Millman and Glass (1967), and by most standard textbooks on experimental design (e.g., Kirk, pp. 208-212, and Winer, pp. 321-332).

In the next two sections we treat the estimation of variance components when both the G and D studies use split-plot designs. Subsequently, we briefly consider the estimation of variance components when the G study is a factorial design and the D study is a split-plot design.

In order to estimate the variance components, we must specify whether the model assumes random, mixed, or fixed effects. The choice among a random, mixed, or fixed effects model is closely related to the choice of a universe of generalization. To treat a facet as a random effect is to say

that the observed conditions of the facet are sampled from an infinite universe of similar conditions. To treat a facet as a fixed effect is to say that the observed conditions of the facet constitute the universe of conditions of the facet. In subsequent sections we discuss the choice among random, mixed, and fixed effects models in the G study, and the implications of this choice for later D studies.

Random Effects Split-Plot Design

The four generalizability coefficients have all been developed in terms of components of variance for a random effects analysis of variance. It was assumed that the classes and the conditions of the two facets were sampled from infinite universes of possible classes and conditions. The formulas for the expected values of the mean squares, based on a random model, are presented in Table 1, where it is assumed that all classes have the same number of students. In Table 1, primes are used with sample sizes in order to distinguish G study sample sizes from subsequent D study sample sizes. Table 1 also provides estimates of the variance components in terms of mean squares from the random effects model.

Insert Table 1 about here

Using the estimated components of variance in Table 1 and the sample sizes from the D study, we can estimate each of the four generalizability coefficients discussed previously. That is, the components of variance from the random effects model can be used to estimate generalizability coefficients for random,

mixed, or fixed effects. Generally, this is the most efficient and useful way to calculate these coefficients. However, if both the G and D study imply the same universe of generalization (e.g., both have items fixed and students random), then one can redefine the structural model and calculate the appropriate generalizability coefficient more directly.

Mixed Effects and Fixed Effects Split-Plot Designs

The analysis of variance in Table 1 treats all effects as random effects. The data from the G study can also be analyzed using a mixed model in which one of the facets is treated as a fixed effect and the other is treated as a random effect. In treating a facet as a fixed effect, the investigator is deciding that his interest is in the observations collected under the finite set of conditions of the fixed facet and in no other possible conditions of that facet. For the split-plot design employing mixed effects, either items or students will be fixed but not both.

If the item facet is fixed and the student facet is random, then the finite set of items under consideration constitute the universe of generalization for the item facet. Thus, when we take the mean over items in Equation 1, the item main effect and all other effects involving items are zero. The resulting structural model for the observed mean score for a class is

$$X_{\underline{c}\cdot\cdot} = \mu + \alpha_{\underline{c}} + \pi_{\cdot(\underline{c})} \cdot \qquad (16)$$

Using Equation 16, the expected observed score variance is

$$\mathcal{E}\sigma^2_{\underline{I}^*}(X_{\underline{c}\cdot\cdot}) = \sigma^2_{\underline{I}^*}(\alpha) + \sigma^2_{\underline{I}^*}(\pi,\alpha\pi)/n_{\underline{s}}; \tag{17}$$

the universe score variance is

$$\sigma^2_{\underline{I}^*}(\nu_{\underline{c}}) = \sigma^2_{\underline{I}^*}(\alpha); \tag{18}$$

and the generalizability coefficient is

$$\mathcal{E}\rho^2(\underline{S},\underline{I}^*) = \frac{\sigma^2_{\underline{I}^*}(\alpha)}{\sigma^2_{\underline{I}^*}(\alpha) + \sigma^2_{\underline{I}^*}(\pi,\alpha\pi)/n_{\underline{s}}}; \tag{19}$$

where the subscript $\underline{I}^*$ is used to indicate that these components of variance are estimated from a mixed model with the item facet fixed. Equations 9 and 19 have the same interpretation, and, as shown below, they are algebraically identical.

Insert Table 2 about here

Table 2 lists the expected values of the mean squares for a mixed model ANOVA with the item effect fixed and all other effects random. Table 2 also provides the estimated values of the variance components in terms of mean squares. In comparing the mixed model in Table 2 with the random model in Table 1, we note that the mean squares in both tables are identical for all sources. Furthermore,

$$\hat{\sigma}^2_{\underline{I}^*}(\alpha) = \frac{\underline{MS}(C) - \underline{MS}(S)}{n'_{\underline{i}}\, n'_{\underline{s}}}$$

$$= \frac{\underline{MS}(C) - \underline{MS}(S) - \underline{MS}(CI) + \underline{MS}(R)}{\underline{n}_i' \underline{n}_s'} + \frac{\underline{MS}(CI) - \underline{MS}(R)}{\underline{n}_i' \underline{n}_s'}$$

$$= \vartheta^2(\alpha) + \vartheta^2(\alpha\beta)/\underline{n}_i' \ . \qquad (20)$$

Similarly, it is straightforward to show that

$$\vartheta^2_{\underline{I}^*}(\pi,\alpha\pi) = \vartheta^2(\pi,\alpha\pi) + \vartheta^2(\beta\pi,\alpha\beta\pi,\underline{e})/\underline{n}_i' \ . \qquad (21)$$

The algebraic equivalence of Equations 9 and 19 is now immediately evident.

Also, if the number of students and the number of items are the same in the G and D study (i.e., $\underline{n}_i' = \underline{n}_i$ and $\underline{n}_s' = \underline{n}_s$), it is easy to show that, for both the random and mixed models, the estimated value of the coefficient is given by

$$\hat{\rho}^2(\underline{S},\underline{I}^*) = \frac{\underline{MS}(C) - \underline{MS}(S)}{\underline{MS}(C)} \ . \qquad (22)$$

The random effects ANOVA outlined in Table 1 attributes the mean square for classes to four sources, two of which involve the sampling variance due to item effects. In the mixed model ANOVA outlined in Table 2, the mean square for classes is attributed to two effects, and the interaction effects involving items do not appear. In the mixed model, there can be no sampling variance for item effects because the mean scores are not based on a sample of items but on the universe of items. The variance that was attributed to item effects in the random

model must now be attributed to the sampling of students or to differences between the universe scores for classes.

A comparison of Table 1 with Table 2 shows that the mean squares used to estimate $\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e})$ and $\sigma^2(\pi,\alpha\pi)$ in the random model are used to estimate $\sigma^2_{\underline{I}*}(\pi,\alpha\pi)$ in the mixed model. Similarly, the mean squares used to estimate $\sigma^2(\alpha\beta)$ and $\sigma^2(\alpha)$ in the random model are used to estimate $\sigma^2_{\underline{I}*}(\alpha)$ in the mixed model. The estimates of the class effect and the student effect are larger for the mixed model than they are for the random model.

This redistribution does not affect the expected observed score variance, but it does change our estimate of the universe score variance. That part of the mean square for classes that is assumed to be due to sampling of items in the random model is assumed to be due to differences between class universe scores in the mixed model.

The mixed model has led to a somewhat simpler expression for $\mathcal{E}\rho^2(\underline{S},\underline{I}*)$, and it will yield the same value for the estimate of this coefficient. Within the assumptions of the mixed model it is not possible to estimate the generalizability coefficient $\mathcal{E}\rho^2(\underline{S},\underline{I})$, which assumes generalization over both facets. For this reason, the mixed model is not recommended for the analysis of the G study data. The mixed model has been introduced mainly to provide additional insight into the nature of the differences between the generalizability coefficients

introduced earlier.

Similar analyses can be carried out for the fixed effects model and the second mixed model in which the student facet is fixed and the item facet is random. For both of these models the generalizability coefficient is equal to the variance attributed to the class effect for the particular model divided by the estimated observed score variance. The numerical estimates of these coefficients will be identical to those previously obtained using components of variance from the random model.

It is generally best to estimate and report components of variance for the random model. If the components of the random model are known, any of the four generalizability coefficients can be estimated; but the components from a model with a fixed facet cannot be used to estimate a generalizability coefficient that assumes generalization over that facet.

Random Effects Factorial Design

If the G study is a factorial design and the D study is a split-plot design, then to calculate generalizability coefficients for the D study we estimate the appropriate variance components from the G study factorial design (see Equation 2). Since the effects are independently sampled, $\sigma^2(\pi,\alpha\pi) = \sigma^2(\pi) + \sigma^2(\alpha\pi)$ in terms of variance components from the factorial design. Similarly, $\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e}) = \sigma^2(\beta\pi) + \sigma^2(\alpha\beta\pi,\underline{e})$ if there is only one observation per class-student-

item combination in the G study; or $\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e}) = \sigma^2(\beta\pi) + \sigma^2(\alpha\beta\pi) + \sigma^2(\underline{e})$ if there are replicated observations. Since $\sigma^2(\alpha)$ and $\sigma^2(\alpha\beta)$ are unconfounded in both the factorial and split-plot design, these variance components have the same interpretation in both designs.

Given the above relationships, we can estimate each of the generalizability coefficients for the split-plot design, with the possible exception of $\mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*)$, Equation 15. This is the appropriate coefficient when the item and student facet are both fixed. It can be estimated only if: (a) there is more than one observation for each class-student-item combination or (b) we assume that both $\sigma^2(\beta\pi)$ and $\sigma^2(\alpha\beta\pi)$ are equal to zero.

Generalizability Coefficients as Expected Values of Correlations

Each of the four generalizability coefficients can also be interpreted as the expected value of a correlation between pairs of measurements on a sample of classes. In order to examine the dependability of class means as correlation coefficients, it is necessary to obtain two measurements on each class. The appropriate procedure for obtaining these two measurements depends on the definition of the universe of generalization.

Generalization over Students and Items

If both students and items are sampled from infinite

universes, each measurement involves a sample of students and a sample of items. The sampling of students and items for a second set of measurements is independent of the first.

Using the random model each measurement has the form

$$X_{\underline{c}\cdot\cdot} = \mu + \alpha_{\underline{c}} + \pi_{\cdot(\underline{c})} + \beta_{\cdot} + \alpha\beta_{\underline{c}\cdot} + \beta\pi_{\cdot\cdot(\underline{c})} + e_{\underline{o}(\underline{c}\cdot\cdot)}. \quad (23)$$

For any pair of measurements, $X_{\underline{c}\cdot\cdot}$ and $X'_{\underline{c}\cdot\cdot}$, each measurement has the same expected observed score variance, given by Equation 3.

The expected value of the covariance of $X_{\underline{c}\cdot\cdot}$ with $X'_{\underline{c}\cdot\cdot}$ is $\sigma^2(\alpha)$. The other effects are sampled independently for the two sets of measurements, and, therefore, the expected values of all other terms in the covariance are zero.

The expected value of the correlation between the two mean scores is approximately equal to the expected value of the covariance divided by the expected value of the variance (Lord & Novick, 1968, pp. 201-203):

$$\mathcal{E}[r(X_{\underline{c}\cdot\cdot}, X'_{\underline{c}\cdot\cdot})] = \frac{\sigma^2(\alpha)}{\sigma^2(\alpha) + \frac{\sigma^2(\pi,\alpha\pi)}{n_{\underline{s}}} + \frac{\sigma^2(\alpha\beta)}{n_{\underline{i}}} + \frac{\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e})}{n_{\underline{i}} n_{\underline{s}}}}, \quad (24)$$

which is identical to $\mathcal{E}\rho^2(\underline{S},\underline{I})$.

Thus, the generalizability coefficient, generalizing over both students and items, is approximately equal to the

expected correlation between two sets of measurements taken on a sample of classes, where the two sets of measurements are based on independent samples of both students and items. A correlation of this kind can be obtained on a set of classes by taking the mean on half of the students and half of the items as one measurement, and the mean on the remaining students and items as the other measurement. Unfortunately, it is not possible to apply the Spearman-Brown formula in this case; consequently, it is generally necessary to use the ANOVA procedures outlined earlier.

Generalization over Students Only

If generalization is to an infinite universe of students and to the finite set of items used in some D study, an appropriate pair of measurements would use the same items but independent samples of students. An estimate of the correlation coefficient could be obtained by taking a random split on each class and correlating the mean scores over the two halves of each class and over all items. In this case the Spearman-Brown formula does apply and can be used to estimate the correlation for full classes.

The derivation of the expected value of the correlation, corrected for class size, over all possible splits on classes is found using the same procedure employed in the previous case. The expected observed score variance for the two measures is again given by Equation 3. The expected value of the covariance,

however, has one additional term. Because the same sample of items is used for both measurements,

$$\underline{\text{cov}}(\alpha\beta_{\underline{c}\cdot}, \alpha\beta'_{\underline{c}\cdot}) = \sigma^2(\alpha\beta)/\underline{n}_{\underline{i}} \ ,$$

and the expected value of the covariance is

$$\sigma^2(\alpha) + \sigma^2(\alpha\beta)/\underline{n}_{\underline{i}} \ .$$

The ratio of the expected covariance to the expected observed score variance is $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$. This coefficient, generalizing over students but not over items, is approximately equal to the expected value of a split-class estimate of reliability that is corrected for class size using the Spearman-Brown formula.

Other Universes of Generalization

Similarly, it can be shown that the coefficient $\mathcal{E}\rho^2(\underline{S}^*,\underline{I})$, for generalization over items but not over students, is approximately equal to the expected value of the split-halves reliability corrected for test length. Also, the generalizability coefficient $\mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*)$, generalizing only over random error, is approximately equal to the correlation between two independent measurements of the class mean using the same items and the same students for both measurements.

Previously Reported Reliability Coefficients

The earlier sections of this paper have presented a discussion of four generalizability coefficients for estimating the dependability of class means. In this section, three coefficients that have been proposed for estimating the reliability of class means will be presented and related to the generalizability coefficients discussed earlier.

In a discussion of the statistical properties of school means, Shaycroft (1962) proposed the following coefficient for estimating the reliability of class means:

$$r_{\overline{AA}} = 1 - \frac{\sigma_A^2}{\underline{n}_{\underline{s}}\sigma_{\overline{A}}^2}(1 - r_{AA}), \tag{25}$$

where

$r_{\overline{AA}} \equiv$ reliability of class means,

$r_{AA} \equiv$ reliability of student scores,

$\underline{n}_{\underline{s}} \equiv$ number of students per class,

$\sigma_{\overline{A}} \equiv$ standard deviation of class means, and

$\sigma_A \equiv$ standard deviation of student scores.

The translation of this formula into the notaticn used in this paper is straightforward.

From previous results, $\sigma_{\overline{A}}^2$ is by definition the expected

observed score variance, $\mathcal{E}\rho^2(X_{\underline{c}\cdot\cdot})$, in Equation 3. Brennan (in press) provides formulas for σ_A^2 and r_{AA} in terms of components of variance from the split-plot model:

$$\sigma_A^2 = \sigma^2(\pi,\alpha\pi) + \underline{K}\sigma^2(\alpha) + [\underline{K}\sigma^2(\alpha\beta) + \sigma^2(\beta\pi,\alpha\beta\pi,\underline{e})]/\underline{n}_{\underline{i}} \quad (26)$$

and

$$r_{AA} = \frac{\sigma^2(\pi,\alpha\pi) + \underline{K}\sigma^2(\alpha)}{\sigma_A^2}, \quad (27)$$

where

$$\underline{K} = \frac{\underline{n}_{\underline{s}}(\underline{n}_{\underline{c}} - 1)}{\underline{n}_{\underline{s}}\underline{n}_{\underline{c}} - 1}. \quad (28)$$

Shaycroft's formula assumes that the G study and the D study use the same data; therefore, there is no need to use primes to distinguish G study and D study sample sizes.

Substituting for r_{AA}, $\sigma_{\bar{A}}^2$, and σ_A^2 in Equation 25 gives Shaycroft's formula in terms of components of variance for the random effects split-plot model:

$$r_{\overline{AA}} = \frac{\sigma^2(\alpha) + \dfrac{\sigma^2(\pi,\alpha\pi)}{\underline{n}_{\underline{s}}} + \underline{L}\left[\dfrac{\sigma^2(\alpha\beta)}{\underline{n}_{\underline{i}}}\right]}{\sigma^2(\alpha) + \dfrac{\sigma^2(\pi,\alpha\pi)}{\underline{n}_{\underline{s}}} + \dfrac{\sigma^2(\alpha\beta)}{\underline{n}_{\underline{i}}} + \dfrac{\sigma^2(\beta\pi,\alpha\beta\pi,\underline{e})}{\underline{n}_{\underline{s}}\underline{n}_{\underline{i}}}}, \quad (29)$$

where

$$\underline{L} = \frac{\underline{n}_{\underline{C}}(\underline{n}_{\underline{S}} - 1)}{\underline{n}_{\underline{S}}\underline{n}_{\underline{C}} - 1}. \qquad (30)$$

The coefficient $\underline{L}$ in Equation 29 depends upon the number of classes and the number of students per class used to estimate r_{AA} and σ_A^2. The coefficients $\underline{K}$ and $\underline{L}$ arise because the student effect is confounded with the class by student interaction in the split-plot design; thus, these coefficients reflect complexities in calculating the appropriate number of degrees of freedom when the sampling of students is stratified by class rather than completely random.

Since $\underline{n}_{\underline{S}}$ and $\underline{n}_{\underline{C}}$ are both greater than one in the split-plot design, $\underline{L}$ is between zero and one; and, therefore,

$$\mathcal{E}\rho^2(\underline{S},\underline{I}^*) <= r_{\overline{AA}} <= \mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*). \qquad (31)$$

Thus, Shaycroft's coefficient (Equation 25) overestimates $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$ and undersetimates $\mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*)$.

In most paractical situations, $\underline{n}_{\underline{s}}$ (and possibly $\underline{n}_{\underline{c}}$) is likely to be fairly large; and, consequently, the coefficient $\underline{L}$ will be close to unity. Assuming that $\sigma^2(\beta\pi,\alpha\beta\pi)$ is close to zero, it follows that r_{AA} will be approximately equal to $\mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*)$.

Wiley (1970) has proposed an intraclass correlation coefficient for estimating the reliability of class means. In his analysis, the estimated universe score variance is $\sigma^2(\alpha) + [\sigma^2(\alpha\beta)]/\underline{n}_{\underline{i}}$, and his coefficient is equivalent to Equation 22, $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$ in the special case where the G and D studies are identical.

Thrash and Porter (1974) have discussed two procedures for estimating the reliability of class means. The first of these procedures is to split each class into two random halves, calculate the correlation between the mean scores for the half-classes, and then use the Spearman-Brown formula to obtain the coefficient for full classes. It has already been shown that the expected value of coefficients calculated in this way, over all possible splits on classes, is given by Equation 22, $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$, which is equivalent to Wiley's coefficient.

The second procedure discussed by Thrash and Porter is to randomly split the test into two halves, correlate the half-test

means for full classes, and then use the Spearman-Brown formula to obtain the coefficient for the full-length test. The expected value of this coefficient, over all random splits on the test, is $\mathcal{E}\rho^2(\underline{S}^*,\underline{I})$. This procedure is implicitly generalizing over items but not over students. Because Thrash and Porter recommend the split-test procedure over the split-class procedure, we will refer to the split-test coefficient as Thrash and Porter's coefficient.

Of the four generalizability coefficients discussed earlier, three are directly related to coefficients that have been proposed for estimating the reliability of class means. The authors are not aware of any analysis of the dependability of class means that uses traditional reliability theory and develops a reliability estimate equivalent to $\mathcal{E}\rho^2(\underline{S},\underline{I})$.

The omission of $\mathcal{E}\rho^2(\underline{S},\underline{I})$ is not due to chance. Traditional reliability theory incorporates a univariate interpretation of error. The assumptions made about error variance differ somewhat, but the errors are always assumed to be drawn from some univariate distribution. Since $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$, $\mathcal{E}\rho^2(\underline{S}^*,\underline{I})$, and $\mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*)$ all arise in the context of models where error is univariate, these coefficients are perfectly compatible with the framework of classical reliability theory. For $\mathcal{E}\rho^2(\underline{S},\underline{I})$, however, the appropriate model involves two distinct components of error whose separate contributions cannot be combined into

a single univariate error term. Therefore, the appropriate model for $\mathcal{E}\rho^2(\underline{S},\underline{I})$ does not arise naturally within classical reliability theory.

Choice of Coefficient when Class is the Unit of Analysis

The choice of an appropriate generalizability coefficient for a particular study depends upon the universe of generalization that is intended.

When class is the unit of analysis, it is difficult to conceive of situations in which the interpretation of the results of a research or evaluation study applies only to the students involved in the study. If the results of studies involving new curricula, teaching techniques, human learning, etc. are to have more than anecdotal interest, they must be generalizable to some universe of students beyond those who actually experienced the treatment under study. The intention to generalize to some larger universe of students is quite explicit whenever variation among students is used to estimate sampling error.

Also, it is usually inappropriate to restrict generalization over items to the particular finite set of items used in some study. However, in educational research and evaluation, it does sometimes happen that the set of items in the study exhaust the universe of behaviors that are of interest. In such cases, it is not appropriate to generalize to a wider

universe of items. For example, if a dental hygiene program is intended to train children in the use of a few basic skills, then the items used to measure the effectiveness of the program might exhaust the universe of interest.

The above observations imply that, for describing the dependability of class means, $\mathcal{E}\rho^2(\underline{S},\underline{I})$ is usually the most appropriate of the four generalizability coefficients discussed in this paper. $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$ appears to be appropriate in some cases; but $\mathcal{E}\rho^2(\underline{S}^*,\underline{I})$ and $\mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*)$ are seldom, if ever, appropriate. From this rationale we conclude that: (a) Wiley's coefficient, which is equivalent to $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$, is appropriate in some cases; (b) Shaycroft's coefficient, which is an upper bound for $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$ and a lower bound for $\mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*)$, is perhaps appropriate in some cases; and (c) Thrash and Porter's coefficient is not likely to be appropriate unless one can make a strong argument for restricting generalization over the student facet.

In summary, clearly there is no universally best coefficient; the most appropriate coefficient can be identified only in the context of a particular study. However, we believe that $\mathcal{E}\rho^2(\underline{S},\underline{I})$ is, in most cases, the most appropriate coefficient to use. We also note that, from examination of Equations 6, 9, 12, and 15, the following relationships hold:

$$\mathcal{E}\rho^2(\underline{S},\underline{I}) <= \mathcal{E}\rho^2(\underline{S},\underline{I}^*) <= \mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*), \tag{32}$$

$$\mathcal{E}\rho^2(\underline{S},\underline{I}) <= \mathcal{E}\rho^2(\underline{S}^*,\underline{I}) <= \mathcal{E}\rho^2(\underline{S}^*,\underline{I}^*), \tag{33}$$

and $\mathcal{E}\rho^2(\underline{S},\underline{I}^*)$ is greater than $\mathcal{E}\rho^2(\underline{S}^*,\underline{I})$ if

$$\frac{n_{\underline{s}}\sigma^2(\alpha\beta)}{n_{\underline{i}}\sigma^2(\pi,\alpha\pi)} > 1. \tag{34}$$

Summary and Conclusions

Using generalizability theory in the context of a split-plot design we have developed and discussed four generalizability coefficients for describing the dependability of class means. We have shown that these coefficients can be obtained in three ways: (a) using variance components from a random effects analysis of variance; (b) using variance components from a mixed or fixed effects analysis of variance; and (c) calculating the expected value of particular correlation coefficients. These four generalizability coefficients have been compared to three previously reported statistics for estimating the reliability of class means. Confusion tends to arise because these reliability coefficients are characterized by different definitions of error. Furthermore, none of these three reliability coefficients is equivalent to the generalizability coefficient $\mathcal{E}\rho^2(\underline{S},\underline{I})$, which is, in our judgment,

generally the most appropriate coefficient.

It is understandable that $\mathcal{E}\rho^2(\underline{S},\underline{I})$ has not been given much attention as a coefficient for describing the dependability of class means. The three previously reported reliability coefficients were developed using a univariate conception of error consistent with classical reliability theory. $\mathcal{E}\rho^2(\underline{S},\underline{I})$, however, depends upon a multivariate conception of error, which is not easily accommodated in classical reliability theory, but arises naturally in generalizability theory.

The generalizability coefficients developed here are descriptive statistics and do not depend upon any parametric assumptions about the distribution of errors. Such parametric assumptions need to be made if one wants to establish confidence intervals or perform statistical tests of significance. However, the advisability of performing such tests of significance is questionable. Even if an estimated variance component does not possess statistical significance, it is an unbiased estimate. As such, it is better to use it than to replace it by zero (Cronbach et al., 1972, pp. 192-193).

In this paper we have considered class as the unit of analysis in a split-plot design. That is, we have used the word "class" to indicate an aggregate unit of analysis with one level of nesting. The extension to multiple levels of nesting

is a relatively straightforward application of the procedures discussed here (see Cronbach et al., 1972).

We have assumed throughout this paper that classes are a random effect. In our judgment, this assumption is generally valid. Nevertheless, the formulas for the four generalizability coefficients from the split-plot design are unchanged if we assume that classes are a fixed effect.

Also, in order to simplify the discussion, we have assumed an orthogonal split-plot design in which the number of students within class, n_s, is constant over all classes. Procedures for doing an analysis of variance for a split-plot design with unequal n's are available in most standard experimental design texts (e.g., Kirk, 1968, pp. 276-282; Winer, 1971, pp. 599-603).

Finally, we note the following recommendation from the most recent edition of Standards for Educational & Psychological Tests (APA, 1974): the "estimation of clearly labeled components of score variance is the most informative outcome of a reliability study, both for the test developer wishing to improve the reliability of his instrument and for the user desiring to interpret test scores with maximum understanding" (p. 49). This is equally true whether the unit of analysis is a person or an aggregate of persons, such as a class. If components of variance from a random effects G study are

reported, then a number of generalizability (or reliability) coefficients are easily estimated, and a single generalizability study can replace a number of separate reliability studies.

References

Abt Associates. Education as experimentation: Evaluation of the Follow Through planned variation model (2 vols.). Cambridge, Mass.: Abt Associates, March 1974.

American Psychological Association. Standards for educational & psychological tests (rev. ed.). Washington, D.C.: American Psychological Association, 1974.

Brennan, R. L. The calculation of reliability from a split-plot factorial design. To be published in Educational and Psychological Measurement, December 1975.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L. J., Ikeda, M., & Avner, R. A. Intraclass correlation as an approximation to the coefficient of generalizability. Psychological Reports, 1964, 15, 727-736.

Haney, W. The dependability of group mean scores. Unpublished special qualifying paper, Harvard Graduate School of Education, October 1974. (a)

Haney, W. Units of analysis issues in the evaluation of project Follow Through. Cambridge, Mass.: The Huron Institute, September 1974. (b)

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. Student evaluations of teaching: The generalizability of class means. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1974.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, Cal.: Wadsworth, 1968

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968

Millman, J., & Glass, G. V. Rules of thumb for the ANOVA table. Journal of Educational Measurement, 1967, 4, 41-51.

Shaycroft, M. F. The statistical characteristics of school means. In J. C. Flanagan, J. T. Dailey, M. F. Shaycroft, D. B. Orr & I. Goldberg, Studies of the American high school. Pittsburgh: University of Pittsburgh, 1962.

Smith, M. S., & Bissell, J. S. Report analysis: The impact of Head Start. Harvard Educational Review, 1970, 40(1), 51-104.

Thrash, S. K., & Porter, A. C. Invalidity of a current method for estimating reliability. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1974.

Wiley, D. E. Design and analysis of evaluation studies. In D. E. Wiley & M. C. Wittrock (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart, and Winston, 1970.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: Mcgraw-Hill, 1971.

Footnotes

[1]Since the observed score ($X_{c\cdot\cdot}$) is the mean over n_i items and n_s students, the contributions of the various effects to the observed score variance (Equation 3) are reduced in accordance with the Central Limit Theorem.

Table 1

Split-Plot Design Analysis of Variance

All Effects Random

Source	df	Expected mean square (MS)
Classes (C)	$n'_c - 1$	$\sigma^2(\beta\pi,\alpha\beta\pi,e) + n'_i\sigma^2(\pi,\alpha\pi) + n'_s\sigma^2(\alpha\beta) + n'_i n'_s\sigma^2(\alpha)$
Students (S)	$n'_c(n'_s - 1)$	$\sigma^2(\beta\pi,\alpha\beta\pi,e) + n'_i\sigma^2(\pi,\alpha\pi)$
Items (I)	$n'_i - 1$	$\sigma^2(\beta\pi,\alpha\beta\pi,e) + n'_s\sigma^2(\alpha\beta) + n'_c n'_s\sigma^2(\beta)$
Classes × Items (CI)	$(n'_c - 1)(n'_i - 1)$	$\sigma^2(\beta\pi,\alpha\beta\pi,e) + n'_s\sigma^2(\alpha\beta)$
Residual (R)	$n'_c(n'_i - 1)(n'_s - 1)$	$\sigma^2(\beta\pi,\alpha\beta\pi,e)$

$$\hat{\sigma}^2(\beta\pi,\alpha\beta\pi,e) = MS(R)$$

$$\hat{\sigma}^2(\alpha\beta) = [MS(CI) - MS(R)]/n'_s$$

$$\hat{\sigma}^2(\pi,\alpha\pi) = [MS(S) - MS(R)]/n'_i$$

$$\hat{\sigma}^2(\alpha) = [MS(C) - MS(S) - MS(CI) + MS(R)]/n'_s n'_i$$

$$\hat{\sigma}^2(\beta) = [MS(I) - MS(CI)]/n'_c n'_s$$

Table 2

Split-Plot Design Analysis of Variance

Classes and Students Random; Items Fixed

Source	df	Expected mean square[a] (MS)
Classes (C)	$n'_c - 1$	$n'_i\sigma^2_{I*}(\pi,\alpha\pi) + n'_s n'_i \sigma^2_{I*}(\alpha)$
Students (S)	$n'_c(n'_s - 1)$	$n'_i\sigma^2_{I*}(\pi,\alpha\pi)$
Items (I)	$n'_i - 1$	$\sigma^2_{I*}(b\pi,ab\pi,e) + n'_s\sigma^2_{I*}(ab) + n'_s n'_i \sigma^2_{I*}(b)$
Classes × Items (CI)	$(n'_c - 1)(n'_i - 1)$	$\sigma^2_{I*}(b\pi,ab\pi,e) + n'_s\sigma^2_{I*}(ab)$
Residual (R)	$n'_c(n'_i - 1)(n'_s - 1)$	$\sigma^2_{I*}(b\pi,ab\pi,e)$

$$\delta^2_{I*}(b\pi,ab\pi,e) = MS(R)$$

$$\delta^2_{I*}(ab) = [MS(CI) - MS(R)]/n'_s$$

$$\delta^2_{I*}(\pi,\alpha\pi) = MS(S)/n'_i$$

$$\delta^2_{I*}(\alpha) = [MS(C) - MS(S)]/n'_s n'_i$$

$$\delta^2_{I*}(b) = [MS(I) - MS(CI)]/n'_s n'_i$$

[a]Greek letters and e indicate random effects; unitalicized Latin letters indicate fixed effects.